AD_____

Award Number: DAMD17-03-1-0429

TITLE: Proteomic Analysis to Identify Novel Circulating Breast Cancer Markers

PRINCIPAL INVESTIGATOR: Francisco J. Esteva, M.D., Ph.D.

CONTRACTING ORGANIZATION: University of Texas MD Anderson Cancer Center
Houston TX 77030

REPORT DATE: June 2006

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
01-06-2006	Final	27 May 03 – 26 May 06

4. TITLE AND SUBTITLE
Proteomic Analysis to Identify Novel Circulating Breast Cancer Markers

5a. CONTRACT NUMBER

5b. GRANT NUMBER
DAMD17-03-1-0429

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Francisco J. Esteva, M.D., Ph.D.

E-Mail: festeva@mdanderson.org

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Texas MD Anderson Cancer Center
Houston TX 77030

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Serum protein profiling using mass spectrometry is a promising approach to identify novel circulating breast cancer markers. In this study, serum was fractionated to deplete highly abundant proteins. After protein digestion, we used liquid chromatography mass spectrometry (LC-MS) to develop diagnostic fingerprints using bioinformatic techniques. Samples were randomized prior to fractionation and mass spectrometry testing. Each fraction was digested with trypsin and subsequently analyzed by LC-MS. Peptides were targeted based on the disease to control peak intensity ratios measured in the averages of all mass spectra in each group and t-tests of the intensity of each individual peak. A series of preprocessing steps (spectral alignment, baseline subtraction, normalization) were employed to produce an expansive list of peptides for further investigation and sequencing. The antibody columns removed 12 of the most abundant proteins in serum. Using LC-MS and bioinformatic analysis we found 17-36 differentially expressed peaks in the Cancer vs. Healthy groups. Efforts are ongoing to identify targeted peptide ion signals using tandem matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS/MS). Serum fractionation using specific antibody columns followed by LC-MS and bioinformatic analysis is a feasible approach to peptide profiling in healthy women and breast cancer patients.

15. SUBJECT TERMS
Breast neoplasms, proteomics, serum, tumor markers biological.

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			USAMRMC
U	U	U	UU	15	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

Introduction

Serum protein profiling using mass spectrometry is a promising approach to identify novel circulating breast cancer markers. One of the major problems with detecting low-abundance proteins in the serum is that they are frequently masked by large, abundant proteins such as albumin and immunoglobulins among others. Therefore, serum protein fractionation is an important consideration. After fractionation, protein profiles can be detected using mass spectrometry. Surface-enhanced laser desorption ionization time-of-flight (SELDI-TOF) has been used to compare protein profiling of serum from healthy individuals and cancer patients. However, SELDI-TOF only yields mass/charge (effectively molecular weight) information and no protein identification. Alternatively, fractionated serum proteins can be analyzed after protease digestion using liquid chromatography mass spectrometry (LC-MS), and the LC-MS profiles can then be compared to develop diagnostic fingerprints using bioinformatic techniques. Differentially regulated peptides can then be identified by MS/MS, allowing verification and antibody-based diagnostics to be developed.

Body

Thirty serum samples from healthy women and breast cancer patients at different stages were fractionated using two separate antibody columns to remove highly abundant proteins. Samples were randomized prior to fractionation and mass spectrometry testing. Briefly, 20 microliters of serum were diluted and injected through a Seppro column and an Agilent column in tandem using appropriate buffers. Each fraction was digested with trypsin and subsequently analyzed by LC-MS.

Serum samples were obtained under protocol LAB02-277 (UTMDACC) with appropriate consent forms on file, aliquoted, and stored frozen at –80. Aliquots (20 ul) from each were separately thawed, diluted 5x in TBS (20 mM pH7.6) and injected onto the depletion columns (Agilent-6, Seppro-12) in tandem flowing at 200 ul per minute in TBS. The effluent was monitored at 280 nm and the flowthrough was collected. The affinity column system was flushed with loading buffer, regenerated with 500 mM Glycine-HCl pH2.0 in TBS and reequilibrated in TBS for the next sample injection. Pilot experiments indicated sample carryover under these conditions was essentially undetectable. The above flowthrough was acetone-precipitated by adding 6 volumes of cold (-20) acetone and standing at –20 overnight. The liquid was carefully decanted, the pellet was washed once with cold (-20) acetone, and the pellet air-dried for several minutes. To this 500 ug trypsin (sequencing grade, Promega) was added in 50 ul 30 mM ammonium bicarbonate and the digestion proceeded for 8 hours at 37C, after which an additional 500 ug trypsin was added and incubated overnight. The digestion was quenched by the addition of acid, and 5 ul injected on the LCMS for profiling.

LCMS was performed using a capillary HPLC (Agilent 1100 capillary) connected to an ESI-TOF mass spectrometer using a nanoflow interface (Mariner, Applied Biosystems). The separation was performed on a 0.150 mm ID x 15 cm C18 reversed-phase column (C18- MS, Grace-Vydac) flowing at 1 uL/min. Samples were injected at 97% A (2% acetonitrile in water containing 0.01% trifluoroacetic acid), and salts flushed out for 40 minutes. Then the mass spectral acquisition was started with the gradient start, proceeding to 50% B (80% acqueous acetonitrile containing 0.01% triflouroacetic acid) over 40 minutes, then ramping up to 90% B over 5 minutes. After flushing at 90% the column was reequilibrated in initial conditions, and two blank gradients were performed to reduce the possibility of peptide carryover into the next run.

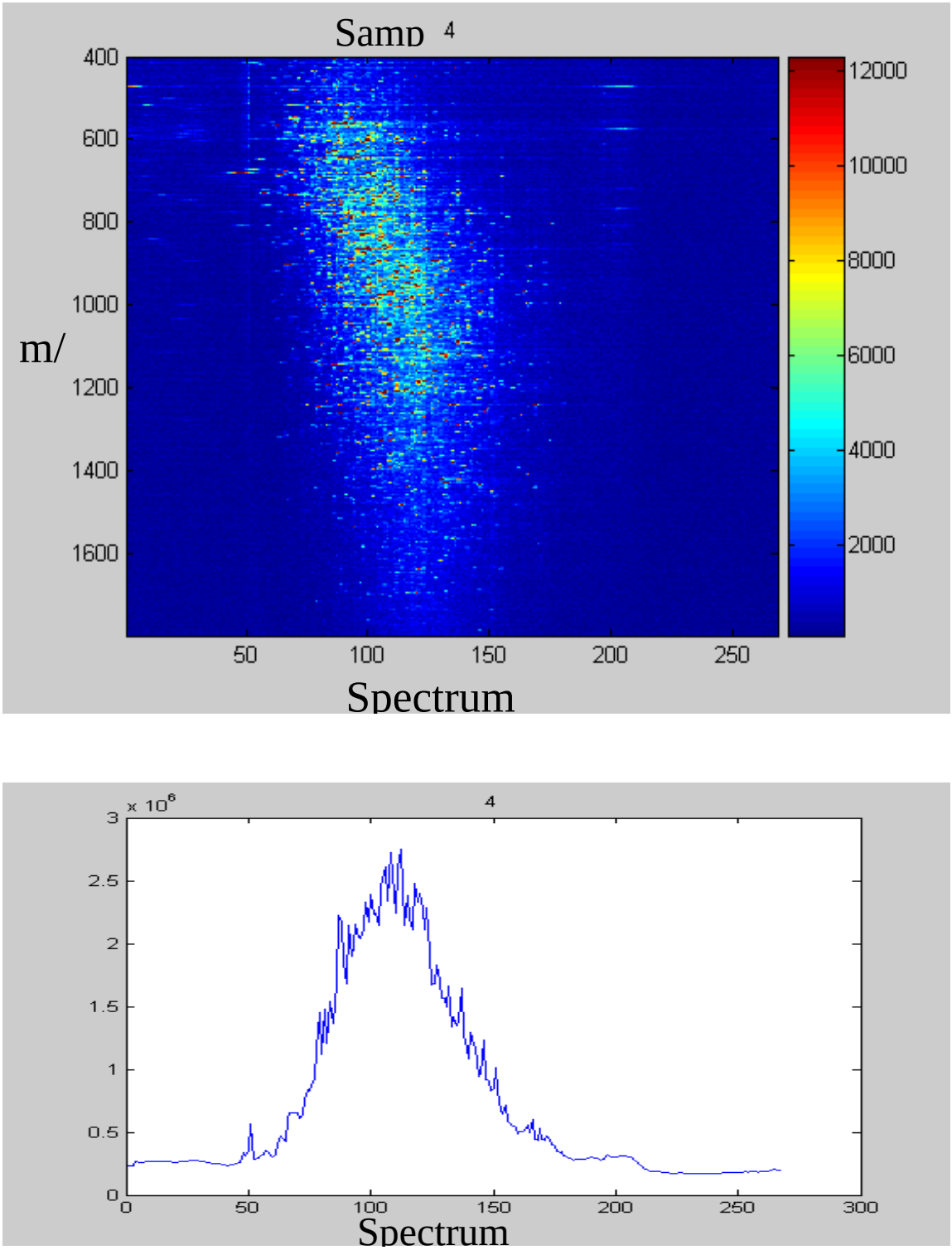


Figure 1. Heat map (upper panel) and total ion chromatogram (lower panel) of one of the samples. Heat map intensity color-code is to the right, TIC ordinate is in counts.

Mass spectra were acquired as the sum of 20 seconds of elution time per spectrum over the course of the 90 minute run, resulting in about 270 spectra per sample. A heat map of the LCMS one of the samples is shown in figure 1 (upper panel), above. The corresponding total ion chromatogram (TIC) is also shown in figure 1 (lower panel). We found there was some variation in the retention times of several major peptide signals, so we adjusted the time coordinates slightly based on apparent retention times of a number of peaks identified as originating from an abundant protein, complement 3. We then calculated the offsets in various regions of the chromatogram, and performed a piece-wise adjustment to the apparent retention times for each run. An example of the adjustment is illustrated in figure 2, below.

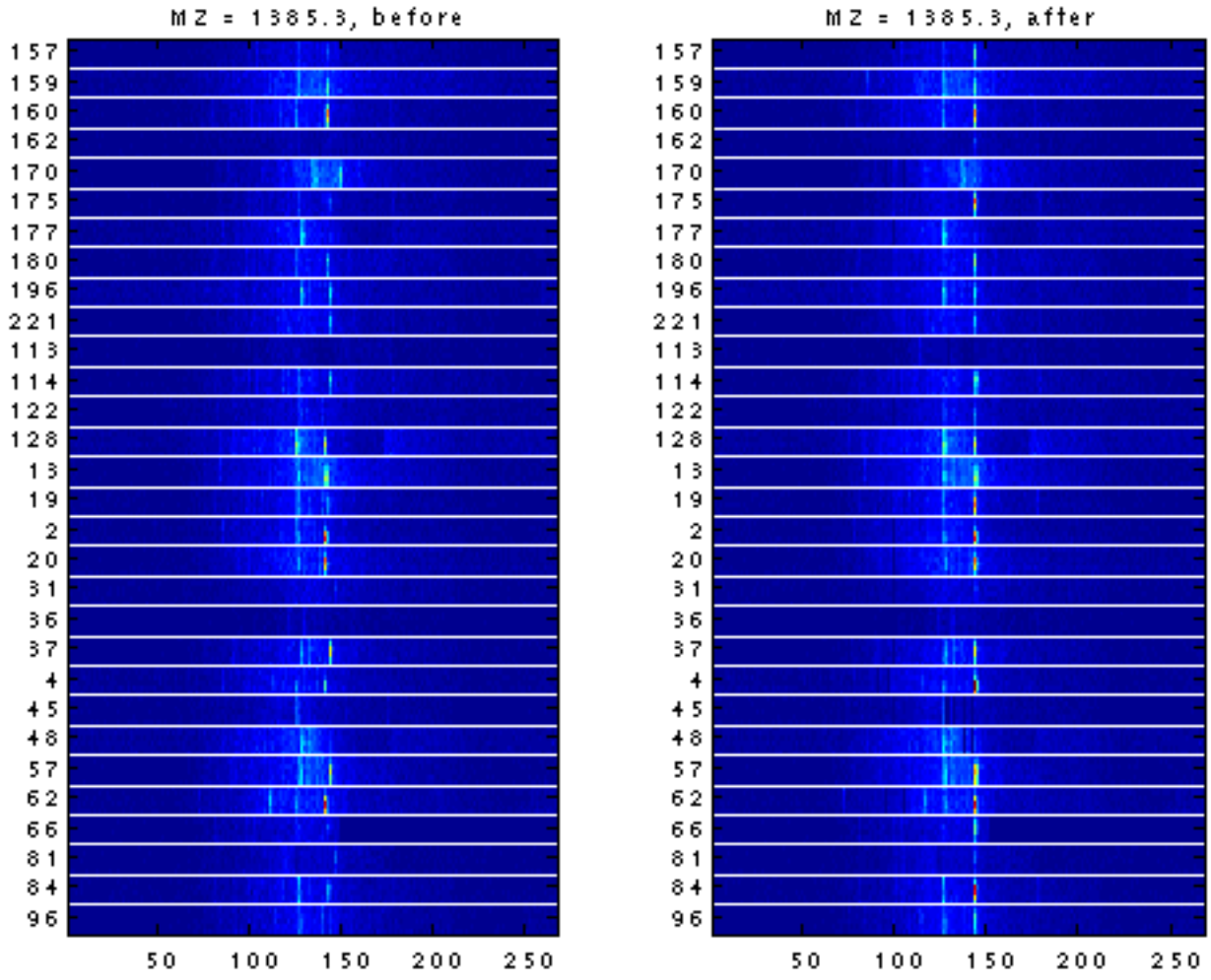


Figure 2. Adjusting the retention time in the neighborhood of the 1385.3 peak. Unadjusted data for mass 1385.3 across the sample set is on the left, the right panel shows the result of the time adjustment.

Our early analysis of these data generated lists of peaks that appeared to be up- or down- regulated based on their t-scores. One of them, expected to have a molecular weight of 1339.7 was found in sample number 48. A portion of this digest was then fractionated and analyzed by LC-MALDI-MS/MS (Dionex-LCPackings HPLC with Probot plate spotting robot, Applied Biosystems 4700 Proteomics Analyzer). Approximately 50 proteins were identified in this experiment with reasonable confidence levels. Of these, one of the proteins found was Protein S. This protein was identified on the basis of a single peptide match, which had the correct MH+ (1340.8) corresponding to the Mr of 1339.7. The match score using the search-engine Mascot was 69, normally a very good score. The spectrum match generated by Mascot is shown in figure 3.

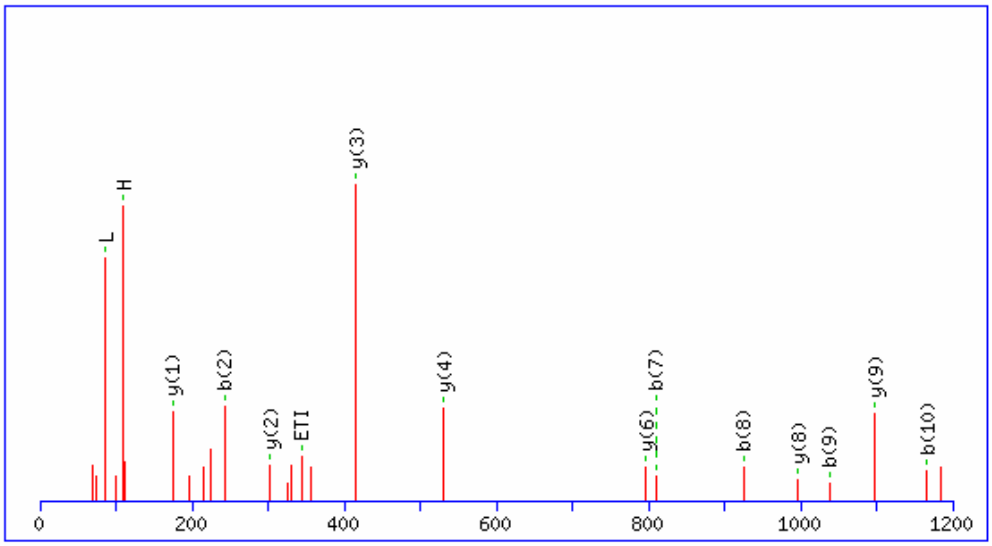
MATRIX SCIENCE Mascot Search Results

Peptide View

MS/MS Fragmentation of IETISHEDLQR
Found in gi|36579, preproprotein S [Homo sapiens]

Match to Query 428 (1340.80,1+) MaldiWellID: 21866, SpectrumID: 65190,

Click mouse within plot area to zoom in by factor of two about that point
Or, Plot from 0 to 1200 Da



Monoisotopic mass of neutral peptide (Mr): 1339.67

Figure 3. Centroided spectrum match output from Mascot for the target peptide at MH+=1340.8. The score for this match was 69.

Identification of Breast Cancer-Specific Markers

Rather than using bioinformatic analysis as a pattern-matching technique, peptides were targeted based on the disease to control peak intensity ratios measured in the averages of all mass spectra in each group and t-tests of the intensity of each individual peak. A series of preprocessing steps were employed to produce an expansive list of peptides for further investigation and sequencing. These steps included spectral alignment, baseline subtraction, normalization, identifying of local maxima, further identifying "large" maxima as peaks, and looking for signs of differential expression.

When the spectra were initially loaded, they were of slightly different offset in elution time. In order to align the spectra correctly, we used 4 known proteins as markers to calibrate the elution times. The markers are at m/z values 421.7,500.8,613.3 and 596.3. We found the elution times for the above peaks for each sample and compared them. The results we got are shown below. These are ordered by their elution times as shown in Table 1.

Table 1. Sample elution time according to the 4 known proteins

Sample_ID	m/z 421.7(2+)	m/z 500.8 (2+)	m/z 613.3(3+)	m/z 596.3(2+)
221	22.8	27.6	32.4	37.5
4	23.4	27.6	32.4	37.2
20	23.7	27.9	32.7	37.5
48	21.9	27.6	33	38.1
159	23.1	27.9	33.3	38.1
19	21.6	27	32.4	37.2
62	21	26.4	31.8	36.9
114	20.4	27.6	33	37.5
196	21.9	27.3	32.7	37.5
***13**	**27.6**	**26.4**	**36.9**	**40.5**
57	21.6	27.3	32.7	37.8
***180**	**28.2**	**27**	**33**	**37.8**
162	22.8	27.6	33	37.8
96	21.6	27.3	32.7	37.8
37	22.5	27.6	33	37.8
170	22.5	27.6	33	37.8
2	21	27	32.4	37.5
31	23.4	29.1	34.5	39
177	22.2	27.3	32.7	37.8
122	21	26.7	32.1	37.2
84	21	27	32.4	37.5
160	21.3	27	32.1	37.5
***113**	**27**	**27.3**	**32.7**	**37.5**
***66**	**26.7**	**27**	**32.4**	**37.5**
***157**	**25.5**	**27.3**	**32.4**	**37.5**
***128**	**27.3**	**27.6**	**33**	**37.8**
***81**	**27.3**	**27.6**	**33**	**37.8**
***175**	**25.8**	**30**	**32.4**	**38.7**
***36**	**23.4**	**28.8**	**33.9**	**36.3**
***45**	**25.8**	**29.7**	**32.4**	**37.2**
Median	22.8	27.45	32.7	37.5

black: healthy
blue: stage 3
red:stage 4

We aligned the sample elution times by using least squares to identify the best linear transformation for getting the observed times to match the median times overall. After refitting the times we interpolated the data onto a common grid of observation times for easier analysis.

Baseline estimating is performed using local minima as "local" peaks and minima are observed at roughly every m/z. Baseline was defined on a per spectrum basis to be the minimum value in a moving window of 1Da and subtracted from every spectrum.

Normalization to the total ion current (TIC) was performed after baseline subtraction.

As a prelude to identifying peaks, we identified local maxima in each spectrum, recorded intensities at the maximum and the index of the maximum within the vector. The maxima were then matched by m/z across spectra; the matching window identified maxima differing in location from the first by less than 0.5 Da. The maxima lists produced had 1448 entries, or roughly one per m/z (the m/z range examined was 400 to 1800). As most of the maxima are due to chemical noise, we wanted to identify only the largest as peaks. To do this, we used a rough analog of the signal to noise ratio. First, we used a sliding window of 90 sequential maxima along the m/z vector and also +/- 20s elution time; the value of 90 was chosen so that more than 2/3 of the maxima in any window were, in our assessment, chemical noise. The 90 maxima values were sorted, and the distance Di between the 30th and 60th value was taken as a robust measure of the spread in maxima heights to be expected in the absence of structure. More than 5 Di in at least two spectra indentified a maximum as a "peak" if it was greater than the median intensity in the 90 maxima window. Having identified peaks (around 3000 here) we next focused on those showing the most differential in their intensities according to T-statistics. We applied two sample t-tests contrasting each of 3 subsets of the disease group (all Cancer, Stage 3, Stage 4) to the control group and flagged peaks with t-values greater than 3, indicating higher intensities in the disease groups.

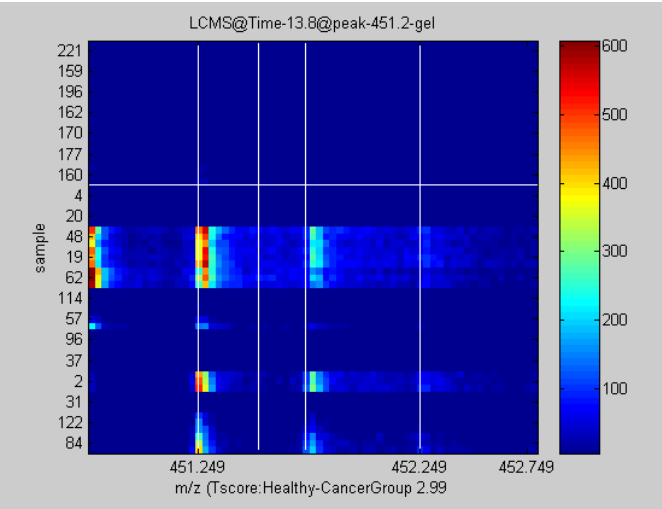
As a quality assessment, we also performed permutation tests. We scrambled the order of the 20 samples randomly choosing the "disease" group and "control" groups. Then we performed T-tests using the key peaks and counted how many had |T|>3, In 1000 random simulations we found 50 interesting peaks on average, which means that in the absence of structure we expect to get about 50 peaks with high t scores. Considering we got around 30 peaks (of these 30 we show only 17 in the table which had strong contrast in the pictures) with higher t-scores, the length of our "interesting" peak list result is not statistically significant enough.

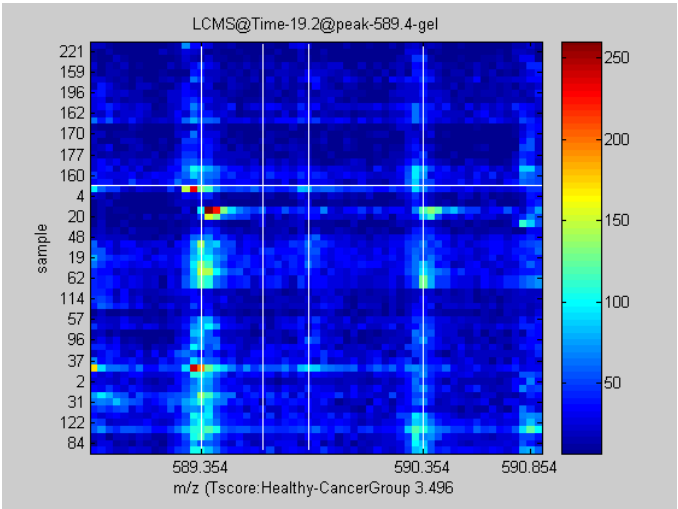
We found 17 "interesting" peaks in the all Cancer vs. Healthy Group contrast as shown in Table 2

Table 2. Cancer Vs Healthy T-test with |T|>3

m/z vector	Time	T-score	Max_intensity
418.2	21.6	3.7369	212
433.3	18.3	3.1329	326
451.2	13.8	2.9905	451
465.2	14.1	3.3331	69
478.3	26.4	2.9762	795
526.3	20.4	3.3092	2453
531.7	21.3	3.3466	816
537.8	21.6	3.036	1068
556.3	12.9	3.4276	43
575.4	21.9	-3.3948	700
576.3	18.3	3.3428	172
587.3	19.8	3.1127	688
592.8	19.5	3.2321	496
671.3	23.1	3.1114	896
688.4	19.5	3.0753	62
779.3	20.7	3.0728	160
1051.5	20.4	3.3132	354

Below are 2 Pictures with strongest contrast



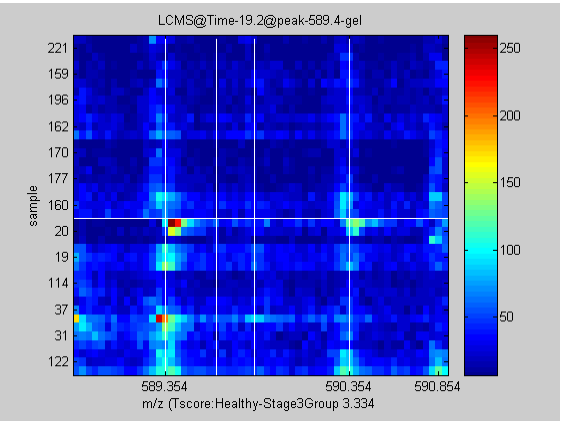


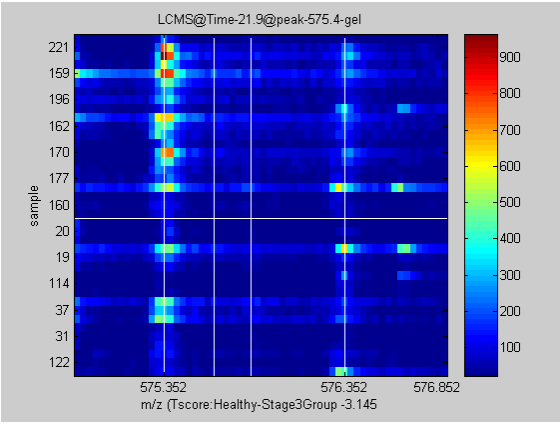
We found 28 interesting peaks in the Stage 3 vs. Healthy Group Contrast

Table 3 Stage 3 Vs Healthy T-test with |T|>3

m/z vector	Time	T-score	Max_intensity
411.3	33.6	3.1611	121
443.3	35.4	2.9542	138
450.7	21.9	-3.1074	752
476.3	24.9	-3.2634	697
489.3	35.1	3.1801	160
508.3	39	2.9598	179
526.3	38.1	2.9845	210
535.3	21.3	3.2102	222
540.3	32.4	3.6654	233
545.3	39.6	2.9663	136
575.4	21.9	-3.1454	700
582.3	42.6	3.2296	105
589.4	19.2	3.334	137
600.3	24.9	-3.8979	216
602.3	24	-3.175	426
617.3	39	3.023	165
624.3	38.1	3.1417	1105
658.4	28.8	-3.2998	495
706.5	39.9	-3.3756	319
736.4	27.3	-3.9796	804
780.4	38.1	3.0632	554
819.9	38.7	-3.1763	389
820.4	38.1	3.4948	343
825.4	37.2	-3.3453	299
877.4	39.9	-3.1311	326
939.4	39.6	-4.1954	174
1128.6	38.7	-3.6662	419
1330.7	39.9	-3.3623	162

Below are 2 Pictures with strongest contrast



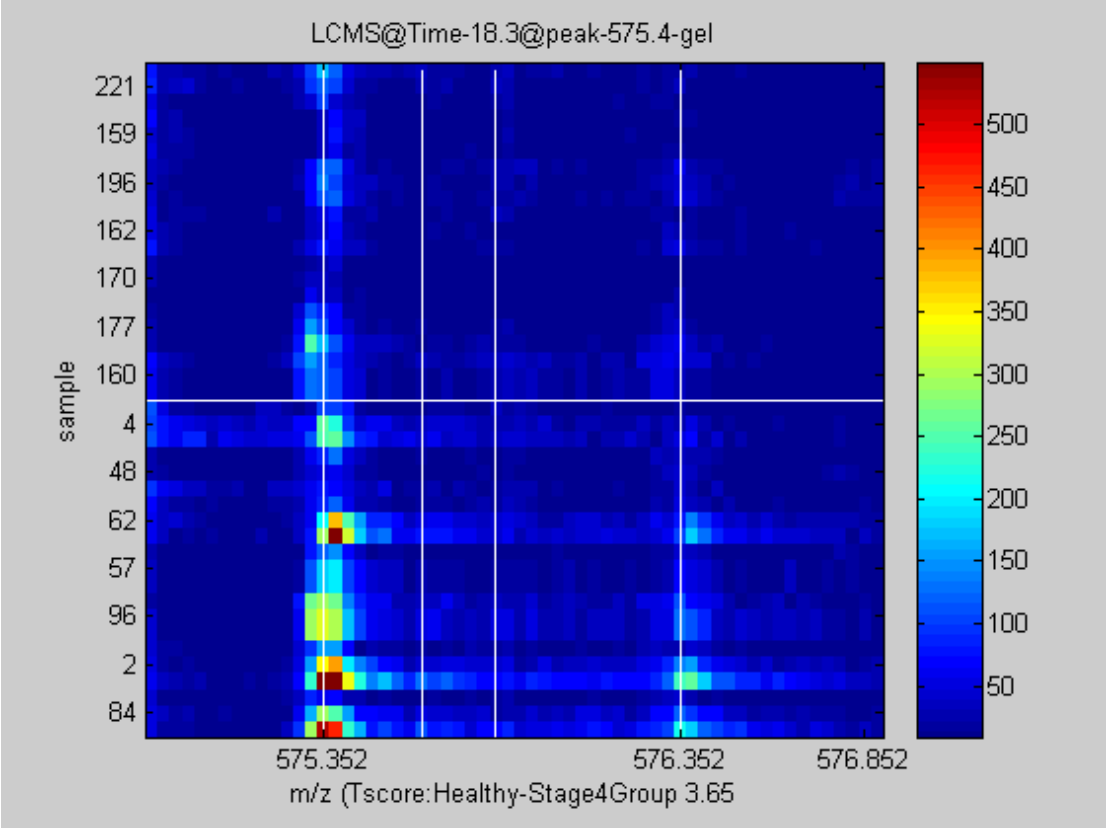


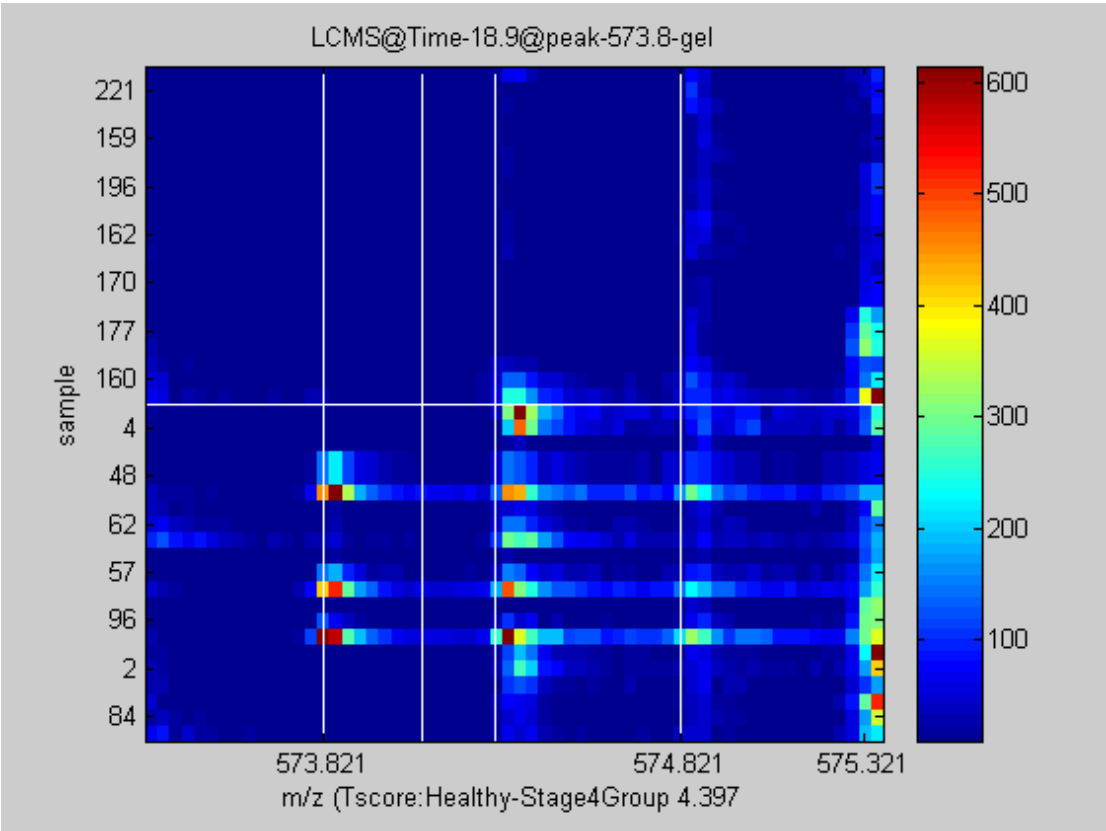
We found 36 interesting peaks in the Stage 4 vs. Healthy group contrast

Table 3 Stage 4 Vs Healthy T-test with T>3

m/z vector	Time	T-score	Max_intensity
404.2	21.3	-3.9416	121
407.3	22.8	-3.604	194
409.2	21.3	-3.8346	89
416.2	23.1	-3.8167	120
418.2	21.3	3.2493	143
426.2	20.4	3.6001	243
451.2	14.1	3.2557	442
469.3	42.6	-3.1279	53
476.3	24.9	-3.3092	697
486.3	21.3	4.8805	250
491.8	20.4	3.3753	235
499.3	24.3	-3.0916	140
505.2	26.4	3.2932	180
526.3	20.4	3.0491	2453
531.7	21.3	3.2793	816
537.8	21.6	3.2657	1068
540.3	16.5	3.2122	70
558.8	22.2	3.2495	153
563.3	24	3.1252	1067
573.8	18.9	4.3966	202
575.4	18.3	3.6502	344
575.4	21.9	-3.4023	700
587.3	19.8	3.6029	439
592.8	19.5	3.5154	496
597.3	28.8	-3.6167	270
610.3	24	4.5385	1027
614.2	23.1	-3.6692	332
684.3	25.8	-3.4952	174
687.4	19.5	3.1566	94
695.3	22.2	3.0432	294
699.3	29.4	-3.2818	357
779.3	23.1	-3.2875	140
787.5	21.6	3.1276	331
797.5	19.5	3.4273	117
874.5	22.5	3.0966	584
1051.5	20.4	3.135	338

Below are 2 Pictures with strongest contrast





Key Research Accomplishments

One of the key accomplishments is the development of novel techniques to assess biomarkers in serum from breast cancer patients. Serum contains millions of peptides and proteins, and fractionation is currently needed to remove highly abundant proteins. Our approach using antibody columns to remove 12 of the most abundant proteins followed by LC-MS is novel, and will likely lead to the identification and validation of breast cancer-specific proteomic signatures in serum.

Further statistical analyses of the dataset presented on this report revealed novel biomarker candidates that will be characterized and validated in future studies.

Reportable Outcomes

The Seppro and Agilent columns removed 20 of the most abundant proteins in serum, including albumin, IgG, Fibrinogen, Transferrin, IgA, IgM, a1-Antitrypsin, Haptoglobin, a1-Acid Glycoprotein, a2-Macroglobulin and HDL (Apolipoproteins A-I and A-II). Using LC-MS and bioinformatic analysis we found 17 differentially expressed peaks in the Cancer vs. Healthy groups; 28 differentially expressed peaks in the Stage 3 vs. Healthy groups; and 36 differentially expressed peaks in the Stage 4 vs. Healthy groups.

Conclusions

Serum fractionation using specific antibody columns followed by LC-MS and bioinformatic analysis may be a feasible approach to peptide profiling in healthy women and breast cancer patients. A key advantage is that detected changes can be identified by ms/ms of the target peptides. A disadvantage compared with the SELDI experiment is that each samples produces about 100 times more data per sample to process.

In summary, serum fractionation using specific antibody columns followed by LC-MS and bioinformatic analysis is a feasible approach to peptide profiling in healthy women and breast cancer patients.

References

Koomen JM, Zhao H, Li D, Nasser W, Hawke DH, Abbruzzese JL, Baggerly KA, Kobayashi R.
Diagnostic protein discovery using liquid chromatography/mass spectrometry for proteolytic peptide targeting.
Rapid Commun Mass Spectrom. 2005;19(12):1624-36.

**Esteva FJ, Zhang B, Hawke D, Zhao H, Baggerly K, Koomen J, Hortobagyi GN, Kobayashi R.
Circulating tumor marker discovery using proteolytic peptide profiling and isotope-coded affinity tags.
Breast Cancer Res Treat 84:2002, 2005.**

Appendices
N/A